BAB II
LANDASAN TEORI

Kanker Payudara

Kanker payudara berasal dari sel-sel lapisan (epitel) saluran susu (85%) atau lobulus (15%) jaringan kelenjar payudara. Awalnya, pertumbuhan kanker terbatas pada saluran atau lobulus ("in situ"), di mana biasanya tidak menimbulkan gejala dan memiliki risiko penyebaran (metastasis) yang minimal. Seiring waktu, kanker dapat terbentuk secara lokal (stadium 0) dan menyerang jaringan payudara di sekitarnya (kanker payudara invasif) dan kemudian menyebar ke kelenjar getah bening terdekat (metastasis regional) atau organ lain di dalam tubuh (metastasis jauh). Ketika seorang wanita meninggal karena kanker payudara, itu karena metastasis yang meluas [9].

Kanker payudara merupakan tipe kanker yang umumya terbentuk di sel-sel payudara dan sel-sel kanker tersebut tumbuh diluar kendali. Kanker payudara dapat terjadi pada semua gender, tetapi kanker ini umumnya lebih sering terjadi pada wanita. Di Australia kanker payudara merupakan tipe kanker yang sangat umum terjadi pada wanita dan kanker paling umum kedua yang menjadi penyebab kematian pada wanita setelah kanker paru-paru [17].

Kanker payudara adalah jenis kanker yang terjadi ketika sel-sel ganas (kanker) berkembang di dalam jaringan payudara. Payudara terdiri dari berbagai jenis sel, termasuk sel-sel kelenjar yang memproduksi susu dan sel-sel penyokong. Kanker payudara umumnya dimulai dari sel-sel kelenjar yang membentuk saluran susu atau lobulus (tempat produksi susu). Kanker payudara dapat berkembang secara perlahan, dimulai dari satu area kecil dalam payudara dan kemudian menyebar ke jaringan di sekitarnya atau ke bagian tubuh yang lebih jauh melalui sistem getah bening atau aliran darah. Hal ini dapat menyebabkan pembentukan benjolan atau

massa yang terasa pada payudara atau perubahan lain pada payudara, seperti perubahan pada kulit atau puting susu.

Kanker payudara adalah penyakit non kulit berbahaya yang paling umum dialami oleh wanita, penyakit tersebut disebabkan oleh beberapa faktor yaitu dari sel dan saluran kelenjar hingga jaringan penopang payudara,kecuali kulit dari payudara. Kanker payudara juga termasuk penyebab nomor dua kematian terbanyak akibat kanker pada wanita setelah kanker serviks, dan cenderung terus meningkat setiap tahunnya [18].

Kanker payudara adalah salah satu jenis kanker yang paling umum di kalangan perempuan, meskipun terjadi juga pada pria tapi jarang. Faktor-faktor risiko yang dapat meningkatkan kemungkinan seseorang mengembangkan kanker payudara antara lain riwayat keluarga dengan kanker payudara, usia, faktor genetik, paparan hormon, obesitas, gaya hidup tidak sehat, dan paparan radiasi [19].

Kanker payudara merupakan salah satu masalah kesehatan yang penting didunia. Kanker payudara merupakan keganasan yang paling sering ditemukan pada wanita diseluruh dunia. Wanita memiliki resiko yang lebih tinggi untuk terkena kanker payudara dibandingkan laki-laki, dikarenakan wanita lebih rentan terhadap hormon *estrogen*. [20].

Data Mining

Data mining adalah proses mengeksplorasi kumpulan data besar yang sebelumnya tidak diketahui [21]. Penambangan pengetahuan juga merupakan bagian dari proses perolehan pengetahuan dari database yang dikenal dengan Knowledge Discovery in Databases (KDD [22]. KDD mewakili penemuan pengetahuan yang terstruktur secara sistematis, mencakup proses standar, dan berkaitan erat dengan manajer dan pengambil keputusan yang aktif. terlibat, dalam proses, berbagi temuan..Perkembangan penting di bidang penambangan data dan penemuan pengetahuan didorong oleh berbagai temuan [23]:

Peningkatan pengumpulan data, seperti yang ditunjukkan oleh pemindaian supermarket, Penyimpanan data di gudang data yang menyediakan akses ke database terkini, yang dapat diandalkan dan Akses yang lebih baik terhadap data melalui penjelajahan Internet dan intranet, Tekanan untuk meningkatkan pangsa pasar dalam perekonomian global, Pengembangan perangkat lunak penambangan data non-komersial, Peningkatan besar dalam pemrosesan dan penyimpanan data fase KDD meliputi pengumpulan data, pengolahan data yang terdiri dari pemilihan data, integrasi data, pembersihan data dan transformasi data [24]. Ada tiga langkah penting dalam KDD sebagai berikut [22]:

1. **Preprocessing Data**

   Tujuan dari proses ini adalah untuk mengubah data masukan ke dalam format yang sesuai untuk analisis lebih lanjut. Langkah ini menggabungkan data dari berbagai sumber, membersihkan data untuk menghilangkan noise dan data duplikat, serta memilih fitur data.

2. ***Data Mining***

   Beberapa teknik penambangan data yang dapat digunakan untuk menangkap pola dan data tersembunyi dalam database antara lain klasifikasi, jaringan syaraf tiruan, pohon keputusan, algoritme genetika, pengelompokan, Pemrosesan Analitik Online (OLAP), dan aturan relasional.

3. **Postprocessing**

   Salah satu tujuan dari proses ini adalah untuk memastikan bahwa hanya hasil yang valid dan berguna yang tersedia bagi pihak yang berkepentingan. Contoh dari proses ini adalah visualisasi, yaitu proses menganalisis dan mengeksplorasi data dan hasil data mining dari berbagai sudut pandang.

Komponen-Komponen Data Mining

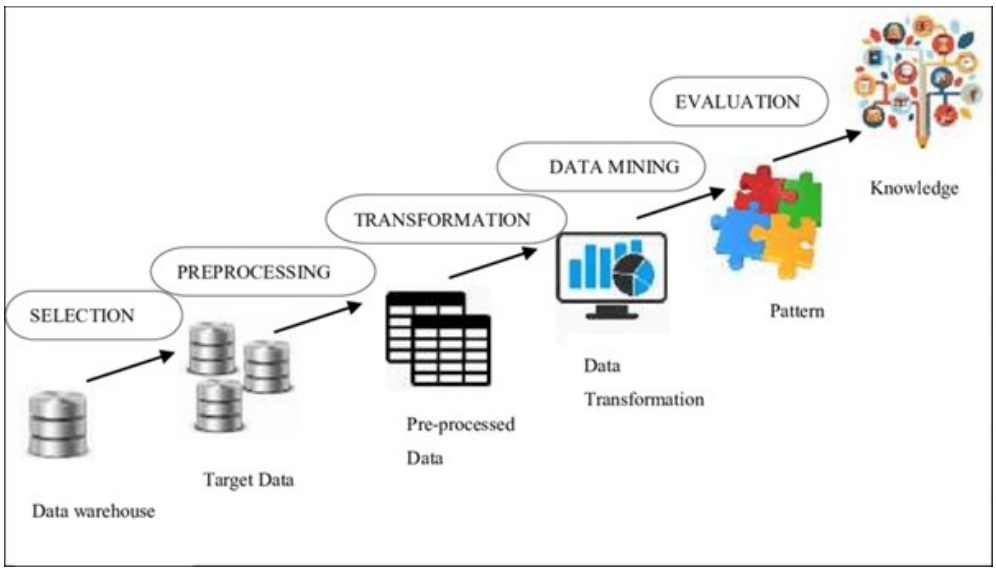
Komponen utama proses klasifikasi adalah sebagai berikut:

1. Kelas, yang merupakan variabel tidak bebas yang berfungsi sebagai label dari hasil klasifikasi;

2. Prediktor, yang merupakan variabel bebas dalam model yang didasarkan pada karakteristik atribut data yang diklasifikasikan, seperti tekanan darah, merokok, minum-minum beralkohol, status perkawinan, dan sebagainya.
3. Set Data Pelatihan adalah sekumpulan data lengkap yang berisi kelas dan prediktor yang dilatih agar model dapat mengelompokkan kelas yang tepat.
4. Set Data Uji adalah sekumpulan data baru yang akan dikelompokkan oleh model untuk mengukur akurasi model yang telah dibuat.

Proses Tahapan Data Mining

Data mining merupakan salah satu dari rangkaian Knowladge Discovery in Database (KDD), KDD merupakan berhubungan dengan teknik integrase dan penemuan ilmiah, interprestasi, dan visualisasi dari pola-pola sejumlah data [23]. Serangkaian proses tersebut memiliki tahapan yang dapat dilihat pada Gambar 2.1



Gambar 2. 1. Tahapan Data Mining

Sumber (https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/)

1. Pembersihan Data (Data Cleaning)

   Proses menghilangkan suara dan data yang tidak konsisten atau tidak relevan dikenal sebagai pembersihan data.
2. Integrasi Data (Data Integration)

   Penggabungan data dari berbagai database baru disebut integrasi data.

3. Seleksi Data (Data Selection)

   Hanya data yang sesuai untuk dianalisis yang akan diambil dari database.
4. Transformasi Data (Data Transformasi)

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam proses penggalian data.
5. Proses Mining

   Merupakan proses utama saat metode diterapkan untuk menemukan penggali yang tepat.
6. Evaluasi Pola (Pattern Evaluation)

   Untuk mengidentifikasi pola-pola menarik kedalam Knowladge Based yang ditentukan.
7. Knowledge, yaitu sebuah hasil yang dicapai berupa pengetahuan atau sebuah informasi.

Fungsi Data Mining

Fungsi data mining dapat dikelompokan ke dalam enam kelompok[23] yaitu :

1. Klasifikasi (classification): menggeneralisasi struktur yang diketahui untuk diaplikasikan pada data-data baru. Misalkan, klasifikasi penyakit ke dalam sejumlah jenis, klasifikasi email ke dalam spam atau bukan.
2. Klasterisasi (clustering): mengelompokkan data, yang tidak diketahui label kelasnya, ke dalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripannya.
3. Regresi (regression): menemukan suatu fungsi yang memodelkan data dengan galat (kesalahan prediksi) seminimal mungkin.
4. Deteksi anomali (anomaly detection): mengidentifikasi data yang tidak umum, bisa berupa outlier,perubahan atau deviasi yang mungkin sangat penting dan perlu investigasi lebih lanjut.
5. Pembelajaran aturan asosiasi (association rule mining) atau pemodelan kebergantungan (dependency modeling): mencari relasi antar variable.

6. Perangkuman (summarization): menyediakan representasi data yang lebih sederhana, meliputi visualisasi dan pembuatan laporan.

Klasifikasi

Klasifikasi merupakan tatanan yang sangat penting dalam komunitas data mining. Klasifikasi adalah teknik penambangan data prediktif yang membuat prediksi tentang nilai data menggunakan hasil yang diketahui dari kumpulan data berbeda. Masalah keakuratan pada banyak algoritme klasifikasi adalah algoritme tersebut dapat mengalami kehilangan informasi jika datanya tidak seimbang, seperti saat distribusi sampel antar kelas sangat miring [25] .

Dalam proses klasifikasi, terdapat variabel target kategoris, seperti braket pendapatan, yang dapat dibagi menjadi tiga kelas atau kategori, seperti berpenghasilan tinggi, menengah, dan rendah. Model data mining meneliti sekumpulan besar catatan, di mana setiap catatan mengandung informasi tentang variabel target serta serangkaian variabel input atau prediktor. Contoh tugas klasifikasi dalam konteks bisnis dan penelitian meliputi:

a. Identifikasi transaksi kartu kredit yang bersifat penipuan
b. Penempatan siswa baru dalam jalur pendidikan yang sesuai dengan kebutuhan khusus mereka
c. Evaluasi aplikasi hipotek untuk menentukan risiko kredit yang terkait
d. Diagnosis penyakit tertentu
e. Verifikasi keaslian surat wasiat, apakah dari almarhum atau dari pihak lain yang mencoba melakukan penipuan
f. Menilai apakah perilaku keuangan atau pribadi tertentu mengindikasikan kemungkinan ancaman teroris.

Klasifikasi yang dilakukan secara manual oleh manusia tanpa bantuan algoritma komputer disebut sebagai klasifikasi manual. Sebaliknya, klasifikasi yang

menggunakan bantuan teknologi melibatkan berbagai algoritma seperti Naïve Bayes, Support Vector Machine, Decision Tree, Fuzzy, dan Neural Networks.[25].

Algoritma K- Nearest Neigbors (KNN

Algoritma K-Nearest Neighbor (K-NN) merupakan salah satu metode yang digunakan untuk mengelompokkan data. Prinsip kerja K-Neares Neighbor (K-NN) adalah mencari jarak terpendek antara data yang akan dievaluasi dengan knn (tetangga) terdekat dalam data latih. Pada fase pelatihan, algoritma hanya menyimpan vektor fitur untuk mengklasifikasikan data pelatihan. Klasifikasi vektor yang sama menghitung pada data pelatihan. Dalam vektor yang sama, kalsifikasi data uji dihitung. Jarak dari vektor terbaru ke semua vektor data pelatihan dihitung, dan sejumlah k terdekat diambil[26]

Decision Tree C-45

Algoritma Decision Tree C 4.5 adalah hasil dari pengembangan algoritma sebelumnya yaitu algoritma *Iterative Dichotomiser* 3 (ID3). Proses pengembangan yang dilakukan pada algoritma C 4.5 menghasilkan beberapa perubahan yang lebih baik diantaranya dapat menangani masalah *missing values* (nilai yang hilang) dalam sebuah dataset, dapat menangani data kontinu (data yang memiliki kemungkinan nilai tidak terbatas) dan melakukan *Prunning*. Algoritma C 4.5 mempunyai input berupa data latih (*training*) serta data sampel [20]. Decision Tree adalah algoritma yang digunakan dalam membuat model keputusan menggunakan struktur pohon atau struktur yang hirarki. Pohon pada decision tree memiliki *root node* dan *node*, *root node* merupakan puncak dari pohon sedangkan *node* merupakan percabangan dari *root node* itu sendiri. Pada setiap *node* decision tree terdapat proses pembuatan keputusan yang menghasilkan dua cabang yaitu “ya” atau “tidak”, pembuatan keputusan sendiri dilakukan dengan menguji suatu variabel, proses pengujian ini terus berlanjut hingga *node* paling bawah atau disebut dengan *leaf node*. Dalam proses decision tree hal yang paling pertama dilakukan adalah memilih *root node*,

salah satu cara dalam pemilihan *root node* ini dapat dilakukan dengan menghitung nilai *gain* pada setiap atribut, sebelum melakukan perhitungan pada nilai *gain* perlu dilakukan perhitungan pada nilai *entropy* terlebih dahulu. Berikut formula *entropy* pada algoritma decision tree [20]:

Formula Entropy:

$$\text{Entropy (S)} = -\sum_{i=1}^{n} Pi * log2(Pi) \quad (1)$$

Keterangan :
n = Jumlah kelas S
P = Proporsi nilai-nilai masuk ke dalam kelas di tingkat Pi

Cross Validation

Metode validasi silang kadang-kadang disebut sebagai metode pengurangan, dimaksudkan untuk mengurangi kuadrat kesalahan prediksi untuk variabel respons, di mana prediksi respons diturunkan dari estimator data [27]. Model atau algoritma akan diajari menggunakan sebagian kecil dari data pelatihan dan akan diuji dengan menggunakan sebagian kecil lagi dari data tersebut, selanjutnya pilihan jenis validasi silang bisa bergantung pada besarnya data yang tersedia. Nilai 10Fold dalam validasi silang merupakan contoh dari metode validasi silang yang digunakan untuk memilih model terbaik, karena cenderung memberikan perkiraan akurasi yang lebih baik dalam proses klasifikasi. Dalam 10Fold Cross Validation, data akan dibagi menjadi 10 bagian dengan ukuran yang sama, sehingga ketika algoritma dinilai, akan menggunakan 10 sub-dataset yang berbeda untuk mengevaluasi performansinya. [20].

Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) adalah algoritma optimasi yang terinspirasi dari perilaku gerak kelompok pada alam semesta, khususnya dalam gerakan kelompok hewan seperti burung, ikan, atau kawanan serangga. Ide dasar di balik PSO adalah mengadaptasi prinsip-prinsip yang mendasari interaksi dan kolaborasi dalam kelompok hewan ke dalam sebuah algoritma komputasional untuk mencari solusi optimal dari suatu permasalahan.

PSO adalah algoritma optimasi stokastik tangguh yang mudah diimplementasikan. Pengaturan parameternya dapat diabaikan. Karena realisasinya yang sederhana dan efisiensi yang tinggi, PSO telah berhasil diterapkan pada berbagai masalah dunia nyata seperti jaringan sensor nirkabel, pemilihan fitur, pengendalian lalu lintas, identifikasi jalan, alokasi tugas, dan pemilihan pengguna kerumunan [[28].

Pada tahun 1995, Kennedy dan Eberhart mengembangkan optimasi gerombolan partikel (PSO) [29]. Dalam PSO, setiap solusi potensial dari permasalahan (disebut sebagai "partikel") dianggap sebagai entitas yang bergerak dalam ruang pencarian. Setiap partikel memiliki posisi dan kecepatan tertentu dalam ruang pencarian yang sesuai dengan solusinya. Konsep utama dari PSO adalah iteratif mengoptimalkan solusi dengan menggerakkan partikel-partikel ini melalui ruang pencarian, dengan harapan bahwa suatu saat akan menemukan solusi terbaik.

Proses pencarian dimulai dengan menginisialisasi sejumlah partikel secara acak dalam ruang pencarian. Setiap partikel memiliki dua komponen penting:

1. Posisi *(Position)*
   Representasi dari solusi yang diusulkan dalam ruang pencarian.
2. Kecepatan *(Velocity)*
   Menentukan seberapa cepat dan ke arah mana partikel akan bergerak dalam ruang pencarian.

Selama iterasi, setiap partikel akan memperbarui posisinya berdasarkan pengalaman pribadinya (Pbest), yaitu posisi terbaik yang pernah dicapai oleh partikel itu sendiri, serta pengalaman bersama (Gbest), yaitu posisi terbaik yang pernah dicapai oleh seluruh partikel dalam populasi. Perubahan posisi dilakukan dengan memperhitungkan kecepatan partikel dan koefisien akselerasi yang mengatur seberapa besar pengaruh pengalaman pribadi dan bersama terhadap perubahan posisi.

Proses iteratif ini berlanjut hingga kriteria berhenti yang ditetapkan tercapai, misalnya jumlah iterasi maksimum tercapai atau solusi yang memenuhi kriteria tertentu telah ditemukan.

Keunggulan dari PSO adalah kemampuannya dalam menemukan solusi optimal dalam ruang pencarian yang kompleks tanpa memerlukan informasi tambahan tentang gradien atau keberlanjutan fungsi objektif. PSO juga relatif mudah diimplementasikan dan dapat diterapkan pada berbagai macam permasalahan optimasi, baik yang stokastik maupun deterministik. Namun, PSO juga memiliki kelemahan, seperti sensitivitas terhadap pemilihan parameter, risiko terjebak dalam optima lokal, dan keterbatasan dalam menangani permasalahan dengan dimensi ruang pencarian yang sangat besar.

Confusion Matrix

Confusion Matrix adalah sebuah teknik yang dipakai untuk mengukur keakuratan atau performa dari data mining. Terdapat empat konsep yang digunakan untuk mewakili hasil klasifikasi, yaitu True Positive (TP) yang mencerminkan observasi yang bernilai positif dan diprediksi positif, False Negative (FN) yang mengindikasikan observasi yang bernilai positif namun diprediksi negatif, True Negative (TN) yang menunjukkan observasi yang bernilai negatif dan diprediksi negatif, serta False Positive (FP) yang menandakan observasi yang bernilai negatif namun diprediksi positif. [30].

Confussion matrix adalah salah satu cara yang sering digunakan pada proses evaluasi model data mining klasifikasi dengan memprediksi kebenaran objek [31]. Matriks konfigurasi adalah tabel yang terdiri dari jumlah baris data uji yang diprediksi benar dan salah dengan model klasifikasi yang digunakan. Tabel Confusion Matrix diperlukan untuk memilih kinerja terbaik dari sebuah model klasifikasi [32]. Confusion matrix diartikan matrix 2x2 yang merepresentasikan hasil dari klasifikasi biner pada suatu dataset. Untuk menghitung performa klasifikasi terdapat beberapa rumus umum yang dapat digunakan. Hasil dari nilai akurasi , presisi dan recall bisa ditampilkan berupa persentase [33].

Accuracy (Akurasi)

Akurasi adalah salah satu matrix yang digunakan untuk mengevaluasi model klasifikasi. Secara informal, akurasi merupakan bagian kecil dari prediksi model kami yang benar. Sedangkan secara formal, akurasi memiliki arti sebagai berikut:

$$\text{Akurasi}=\frac{\text{Number of Correect Prediction}}{\text{Total Number of Prediction}} \quad (2)$$

Akurasi juga dapat dihitung dalam hal negatif dan positif untuk klasifikasi biner sebagai berikut:

$$\text{Akurasi}=\frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Dimana TP = True Positif
TN = True Negatif
FP = False Positif
FN = False Negatif

Precision

Precision dalam Confusion Matrix didefinisikan sebagai rasio item terkait yang dipilih dengan semua item yang dipilih. Akurasi adalah kemungkinan bahwa item

yang dipilih terkait. Dapat diartikan sebagai kecocokan antara permintaan informasi dan respons terhadap permintaan itu [34].

Recall

Recall adalah rasio jumlah dokumen teks terkait yang dikendalikan di antara semua dokumen teks yang relevan dalam suatu koleksi [33]. Recall adalah probabilitas bahwa item terkait dipilih. Recall dapat dihitung sebagai jumlah rekomendasi relevan yang dipilih oleh pengguna dibagi dengan jumlah semua rekomendasi yang relevan, dipilih dan tidak dipilih[34].

Kurva ROC dan AUC

Dalam Machine Learning, pengukuran kinerja adalah tugas penting. Jadi dalam masalah klasifikasi, kita dapat mengandalkan Kurva AUC - ROC. Ketika kita perlu memeriksa atau memvisualisasikan kinerja masalah klasifikasi multi-kelas, kita menggunakan kurva AUC (Area Under The Curve) ROC (Receiver Operating Characteristics). Ini adalah salah satu metrik evaluasi terpenting untuk memeriksa kinerja model klasifikasi apa pun. Itu juga ditulis sebagai AUROC (Area Di Bawah Karakteristik Operasi Penerima) [35]. Metode umum untuk menghitung daerah dibawah kurva ROC yaitu Area Under Curve (AUC) dimana bidang yang berada dibawah kurva mempunyai nilai yang selalu berada pada nilai 0,0 dan 1,0. Namun yang menarik untuk dihitung adalah yang mempunyai luas diatas 0,5, semakin tinggi luasnya maka akan semakin baik seperti yang disajikan berikut ini [32]:

0.9-1.0 = klasifikasi yang sangat baik (*Excellent Classification*)
0.8-0.9 = klasifikasi baik (*Good Classification*)
0.7-0.8 = klasifikasi rata-rata (*Fair Classification*)
0.6-0.7 = klasifikasi rendah (*Poor Classification)*
0.5-0.6 = kegagalan (*Failure Classification)*

Rapidminer

Rapidminer adalah software yang digunakan untuk mengolah data. Dengan pemanfaatan prinsip dan algoritma data mining. Rapidminer mengekstrak pola dari kumpulan data besar dengan menggabungkan metode statistik, kecerdasan buatan, dan basis data. Rapidminer memungkinkan pengguna untuk dengan mudah menghitung sejumlah besar data menggunakan operator [36]. RapidMiner memiliki Graphical User Interface (GUI) yang sangat efektif untuk desain proses analitis. Ini berisi berbagai Repositori untuk proses, operator, data dan membantu dalam manajemen metadata. Ini membantu dalam perbaikan bug dan deteksi kesalahan. Ini adalah alat visualisasi, mudah digunakan tanpa pengkodean dan merupakan paket lengkap dan serbaguna, berisi ratusan pendekatan yang tersedia untuk integrasi data, pembelajaran mesin, dan simulasi [37].

Penelitian Terkait

Metode pembelajaran mesin (*Machine Learning*) telah diterapkan pada data kanker payudara sebagai alat untuk mengklasifikasikan jenis kanker payudara, baik secara mandiri maupun dalam kombinasi dengan data lain untuk validasi atau melengkapi informasi. Beberapa studi yang menggunakan metode klasifikasi ML untuk memprediksi kanker payudara dijelaskan di bawah ini.

Tabel 2. 1. Penelitian Terkait

No	Judul / Penulis / Tahun	Dataset	Metode	Hasil
Study Clasifing Breast Cancer Using Dataset				
1	Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers Fatin Kadhim Nasser*, Suhad Faisal Behadili Iraqi Journal of Science, 2022[11]	Rumah Sakit Pengajaran Onkologi, Medical City, Baghdad, Irak	Pruning Decision Tree dengan kriteria Gini index dan k-Nearest Neighbors (KNN) Algorithm	Akurasi dari model pruning decision tree dengan kriteria Gini index adalah 87,30%

No	Judul / Penulis / Tahun	Dataset	Metode	Hasil
Study Clasifing Breast Cancer Using Dataset				
2	Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer Harikumar Rajaguru, Sannasi Chakravarthy S R 2019.[12]	Wisconsin Diagnostic Breast Cancer (WDBC) dataset yang tersedia di UCI repository	K-Nearest Neighbor (KNN) Decision Tree	K-Nearest Neighbor (KNN) akurasi 95.61% Decision Tree akurasi 91.23%
3	Klasifikasi Penyakit Kangker Payudara Menggunakan Algoritma K-Nearest Neighbor Ramdhani, W., Bona, D., Musyaffa, R., & Rozikin, C. 2022, [38]	Wisconsin Breast Cancer Dataset Kaggle	K-Nearest Neighbor (KNN)	KNN Nilai akurasi 97% dengan nilai k=9
4	Performance Comparison of Machine Learning Techniques for Breast Cancer Detection Penulis: Dada Emmanuel Gbenga *, Ngene Christopher , Daramola Comfort Yetunde Tahun: 2017	Wisconsin Breast Cancer dataset	Radial Based Function (RBF) Network, Support Vector Machine (SVM), Simple Linear Logistic Regression Model (SL), Naïve Bayes (NB), k-Nearest Neighbour (kNN), AdaBoost, Fuzzy Unordered Role Induction algorithm (Fuzzy), and Decision Tree (J48)	Radial Based Function (RBF) Network: 89,4% - Support Vector Machine (SVM): 93,6% - Simple Linear Logistic Regression Model (SL): 92,0% - Naïve Bayes (NB): 92,4% - k-Nearest Neighbour (kNN): 91,3% - AdaBoost: 91,2% - Fuzzy Unordered Role Induction algorithm (Fuzzy): 92,0% - Decision Tree (J48): 90,6%

No	Judul / Penulis / Tahun	Dataset	Metode	Hasil
Study Clasifing Breast Cancer Using Dataset				
5	Komparasi Algoritma Decision Tree, Naive Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara Muhammad Abdul Jabbar, Erfan Hasmin , Sunardi, Cucut Susanto, Wilem Musu 2022	UCI Machine Learning Repository	Decision Tree, Naive Bayes, dan K-Nearest Neighbors	K-Nearest Neighbors 98% pada metode Hold-Out dan 96% pada metode K-Fold, Naive Bayes 95% pada metode Hold-Out dan 95% pada metode K-Fold, dan Decision Tree 94% pada metode Hold-Out dan 93% pada metode K-Fold
6	Komparasi Penerapan Metode Bagging dan Adaboost pada Algoritma C4.5 untuk Prediksi Penyakit Stroke Nur Diana Saputri*, Khalid Khalid, Dwi Rolliawati 2022	UCI Dataset Stroke	Metode Bagging dan Adaboost pada algoritma C4.5	Akurasi C45 tanpa metode bagging & Adabost. 92,87%. C45 Metode Bagging Akurasi 95,02% Metode Adaboost Akurasi 94,63%
7	Komparasi Performansi Algoritma Pengklasifikasi KNN, Bagging Dan Random Forest Untuk Prediksi Kanker Payudara Agung Mulyo Widodo , Nizirwan Anwar , Bambang Irawan , Lista Meria , Andika Wisnujati 2021	UCI Machine Learning Repository Wisconsin Breast Cancer dataset	KNN, Bagging Dan Random Forest	Algoritma Bagging: Akurasi 73,29% Algoritma KNN (IBk): Akurasi 74,37% Algoritma Random Forest: Akurasi 74,37%
8	Penggunaan Algoritma C4.5 Dalam	Dataset Wisconsin	C4.5	Nilai akurasi klasifikasi

No	Judul / Penulis / Tahun	Dataset	Metode	Hasil
Study Clasifing Breast Cancer Using Dataset				
	Menentukan Biopsy Pada Penderita Kanker Payudara Adi Candra P , Bakhtiyar Hadi P	Breast Cancer (WBC) UCI Dataset Repository.		C4.5 94.56% dan nilai AUC untuk algoritma C4.5 0.941.

1. Penelitian yang dilakukan oleh Fatin Kadhim Nasser dan Suhad Faisal Behadili (2022) menggunakan algoritma Pruning Decision Tree dengan kriteria Gini index dan K-Nearest Neighbors (KNN) untuk deteksi kanker payudara. Penelitian ini mencapai akurasi sebesar 87.30% dalam mengklasifikasikan tumor payudara.

2. Penelitian yang dilakukan Harikumar Rajaguru, Sannasi Chakravarthy S R. yang berjudul "Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer." Menunjukkan hasil bahwa algoritma K-Nearest Neighbor (KNN) memiliki akurasi sebesar 95.61% dalam klasifikasi kanker payudara, sedangkan algoritma Decision Tree memiliki akurasi sebesar 91.23% .Hal ini menunjukkan bahwa KNN memberikan hasil yang lebih akurat dalam penelitian ini .

3. Penelitian yang dilakukan oleh Ramdhani, W., Bona, D., Musyaffa, R., & Rozikin, C. penelitian ini bertujuan untuk mengklasifikasikan penyakit kanker payudara menggunakan algoritma K-Nearest Neighbor (KNN). Hasil penelitian menunjukkan bahwa menggunakan algoritma KNN dengan nilai k=9, model yang dikembangkan mencapai tingkat akurasi sebesar 97%. Artinya, model tersebut mampu mengklasifikasikan kasus kanker payudara dengan tingkat keakuratan yang tinggi.

   Penelitian ini memberikan kontribusi dalam pengembangan metode klasifikasi untuk penyakit kanker payudara menggunakan algoritma KNN. Hasil yang

diperoleh menunjukkan potensi penggunaan algoritma ini dalam membantu diagnosis dan deteksi dini kanker payudara.

4. Penelitian yang dilakukan oleh Dada Emmanuel Gbenga, Oladipupo Funke, dan Olatunji Sunday Oyewola menggunakan berbagai teknik pembelajaran mesin untuk deteksi kanker payudara. Dalam studi perbandingan tersebut, hasil yang diperoleh adalah sebagai berikut:
   - Radial Based Function (RBF) Network: 89,4%
   - Support Vector Machine (SVM): 93,6%
   - Simple Linear Logistic Regression Model (SL): 92,0%
   - Naïve Bayes (NB): 92,4%
   - k-Nearest Neighbour (kNN): 91,3%
   - AdaBoost: 91,2%
   - Fuzzy Unordered Role Induction algorithm (Fuzzy): 92,0%
   - Decision Tree (J48): 90,6%

   Hasil penelitian ini menunjukkan bahwa teknik pembelajaran mesin, termasuk algoritma K-Nearest Neighbors (kNN), efektif dalam mengklasifikasikan tumor payudara sebagai jinak atau ganas. Meskipun kNN mencapai tingkat akurasi 91,3%, terdapat beberapa teknik lain yang mencapai tingkat akurasi yang lebih tinggi, seperti SVM dengan akurasi 93,6%. Informasi ini dapat memberikan wawasan penting dalam pengembangan model deteksi kanker payudara yang lebih akurat.

5. Penelitian ini bertujuan untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu Decision Tree, Naive Bayes, dan K-Nearest Neighbors (KNN), dalam konteks klasifikasi kanker payudara. Penelitian ini dilakukan oleh Muhammad Abdul Jabbar, Erfan Hasmin, Sunardi, Cucut Susanto, dan Wilem Musu pada tahun 2022, dengan menggunakan data dari UCI Machine Learning Repository.

   Hasil penelitian menunjukkan bahwa algoritma K-Nearest Neighbors (KNN) mencapai akurasi sebesar 98% pada metode Hold-Out dan 96% pada metode

K-Fold. Algoritma KNN dapat mengklasifikasikan sampel kanker payudara dengan sangat baik, dengan tingkat akurasi yang tinggi pada kedua metode pengujian.

Selanjutnya, algoritma Naive Bayes mencapai akurasi sebesar 95% pada metode Hold-Out dan 95% pada metode K-Fold. Meskipun tingkat akurasi Naive Bayes sedikit lebih rendah dibandingkan dengan KNN, namun tetap menunjukkan kinerja yang baik dalam klasifikasi kanker payudara.

Sementara itu, algoritma Decision Tree mencapai akurasi sebesar 94% pada metode Hold-Out dan 93% pada metode K-Fold. Meskipun algoritma ini memberikan tingkat akurasi yang sedikit lebih rendah dibandingkan dengan algoritma KNN dan Naive Bayes, tetap memberikan hasil yang cukup baik dalam klasifikasi kanker payudara.

Dalam kesimpulan penelitian ini, dapat disimpulkan bahwa ketiga algoritma klasifikasi yang dievaluasi (Decision Tree, Naive Bayes, dan K-Nearest Neighbors) memberikan hasil yang baik dalam klasifikasi kanker payudara. Algoritma KNN memiliki tingkat akurasi tertinggi, diikuti oleh Naive Bayes dan Decision Tree. Namun, keputusan penggunaan algoritma klasifikasi yang paling tepat harus didasarkan pada pertimbangan lebih lanjut seperti kebutuhan aplikasi, interpretabilitas model, dan efisiensi komputasi.

6. Penelitian ini bertujuan untuk membandingkan kinerja dua metode ensambel, yaitu metode Bagging dan Adaboost, dalam penerapan algoritma C4.5 untuk prediksi penyakit stroke. Penelitian ini dilakukan oleh Nur Diana Saputri, Khalid Khalid, dan Dwi Rolliawati pada tahun 2022, menggunakan dataset tentang stroke.

   Hasil penelitian menunjukkan bahwa algoritma C4.5 tanpa metode ensambel mencapai akurasi sebesar 92,87%. Metode Bagging, yang diterapkan pada algoritma C4.5, menghasilkan akurasi sebesar 95,02%. Sementara itu, metode Adaboost, yang juga diterapkan pada algoritma C4.5, mencapai akurasi sebesar 94,63%.

Dari hasil tersebut, dapat disimpulkan bahwa penerapan metode ensambel (Bagging dan Adaboost) pada algoritma C4.5 dapat meningkatkan akurasi prediksi penyakit stroke dibandingkan dengan menggunakan algoritma C4.5 tanpa metode ensambel. Metode Bagging memberikan akurasi tertinggi sebesar 95,02%, diikuti oleh metode Adaboost dengan akurasi 94,63%. Hal ini menunjukkan bahwa kedua metode ensambel tersebut efektif dalam meningkatkan kinerja algoritma C4.5 dalam prediksi penyakit stroke.

Namun, penting untuk dicatat bahwa selain akurasi, faktor lain seperti kecepatan komputasi dan interpretasi model juga perlu dipertimbangkan dalam memilih metode ensambel yang paling sesuai untuk aplikasi prediksi penyakit stroke.

7. Penelitain yang berjudul Komparasi Performansi Algoritma Pengklasifikasi KNN, Bagging Dan Random Forest Untuk Prediksi Kanker Payudara yang dilakukan oleh Agung Mulyo Widodo, Nizirwan Anwar , Bambang Irawan, Lista Meria , Andika Wisnujati pada tahun 2021. Penelitian ini bertujuan untuk membandingkan akurasi tiga algoritma klasifikasi, yaitu KNN, Bagging, dan Random Forest, dalam memprediksi kanker payudara menggunakan dataset kanker payudara.
   Hasil penelitian menunjukkan bahwa algoritma KNN memiliki akurasi tertinggi dibandingkan dengan algoritma lainnya, yaitu Bagging dan Random Forest. Algoritma KNN mampu mengklasifikasikan sampel dengan tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma lainnya.

   Pada pengujian dengan metode 10-fold cross-validation, algoritma KNN menghasilkan 206 kejadian yang diprediksi dengan benar, dari total 277 kejadian yang diuji. Sebaliknya, algoritma KNN juga menghasilkan 71 kejadian yang salah diprediksi. Akurasi dari algoritma KNN adalah 74,37%.

   Hasil penelitian ini menunjukkan bahwa algoritma KNN memiliki performa yang lebih baik dalam memprediksi kanker payudara dibandingkan dengan algoritma Bagging dan Random Forest. Oleh karena itu, algoritma KNN dapat

dianggap sebagai pilihan yang lebih baik dalam penggunaan teknik klasifikasi untuk mendeteksi kanker payudara.

Penelitian ini memberikan kontribusi penting dalam pengembangan sistem pengambilan keputusan berbasis data mining untuk mendukung diagnosis kanker payudara. Dengan menggunakan algoritma KNN, diharapkan dapat meningkatkan akurasi prediksi kanker payudara dan membantu dalam upaya deteksi dini serta pengobatan yang lebih efektif.

8. Penelitian ini bertujuan untuk menggunakan algoritma klasifikasi C4.5 dalam menentukan keputusan biopsi pada penderita kanker payudara. Penelitian ini menggunakan dataset Wisconsin Breast Cancer (WBC) dari UCI Dataset Repository. Hasil penelitian menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi klasifikasi sebesar 94.56% dan nilai AUC (Area Under Curve) sebesar 0.941.

   Algoritma C4.5 merupakan metode klasifikasi berbasis pohon keputusan yang digunakan untuk memprediksi kelas atau keputusan berdasarkan atribut yang ada pada dataset. Pohon keputusan yang dihasilkan oleh algoritma C4.5 digunakan sebagai model untuk melakukan klasifikasi pada data baru. Algoritma ini menggunakan metode information gain untuk memilih atribut yang paling informatif dalam memisahkan data menjadi kelompok yang berbeda.

Penelitian ini mengaplikasikan algoritma C4.5 pada dataset kanker payudara untuk menentukan apakah seorang pasien memerlukan biopsi atau tidak. Hasil evaluasi menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi sebesar 94.56%. Artinya, algoritma ini mampu dengan akurat memprediksi apakah seorang pasien memerlukan biopsi berdasarkan atribut yang ada pada data pasien tersebut.

Selain itu, nilai AUC untuk algoritma C4.5 adalah 0.941. Nilai AUC digunakan sebagai metrik untuk mengukur kualitas klasifikasi. Semakin tinggi nilai AUC, semakin baik kinerja klasifikasi. Dalam kasus ini, nilai AUC yang mendekati 1

menunjukkan bahwa algoritma C4.5 memiliki kemampuan yang baik dalam membedakan pasien yang memerlukan biopsi dan yang tidak.

Dengan hasil akurasi klasifikasi yang tinggi dan nilai AUC yang baik, penelitian ini menunjukkan bahwa algoritma C4.5 dapat digunakan sebagai metode yang efektif dan akurat dalam menentukan keputusan biopsi pada penderita kanker payudara. Penggunaan algoritma ini dapat membantu dokter dan tenaga medis dalam mengambil keputusan yang tepat terkait proses diagnosa dan pengobatan kanker payudara.

Dengan demikian, penelitian-penelitian ini menunjukkan bahwa algoritma KNN dan Decision Tree C4.5 dapat memberikan hasil yang baik dalam deteksi penyakit kanker payudara. Penggunaan algoritma-algoritma ini memberikan harapan untuk meningkatkan diagnosis dini dan pengobatan penyakit kanker payudara.

Penelitian tentang kanker payudara masih terus dilakukan karena masih menjadi masalah besar baik di dunia maupun di Indonesia. Tabel 2.1 merangkum penelitian-penelitian tersebut, termasuk dataset, metode klasifikasi ML (*Machine Learning*), yang digunakan, dan tingkat akurasi yang dicapai.